National South African HIV prevalence estimates robust despite substantial test non-participation

G Harling,[1,2] MA, MPH, ScD; **S Moyo**,[3] MB ChB, MPH, PhD; **M E McGovern**,[4,5] PhD; **M Mabaso**,[3] MSc, PhD; **G Marra**,[6] MSc, PhD; **T Bärnighausen**,[2,7,8] MD, MSc, MSc, ScD; **T Rehle**,[3,9] MD, PhD

[1] *Research Department of Infection and Population Health, Institute for Global Health, University College London, UK*
[2] *Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, USA*
[3] *Human Sciences Research Council, Cape Town, South Africa*
[4] *Centre for Health Research at the Management School (CHaRMS), Queen's University Belfast, UK*
[5] *UKCRC Centre of Excellence for Public Health (Northern Ireland), UK*
[6] *Department of Statistics, University College London, UK*
[7] *Institute of Public Health, University of Heidelberg, Germany*
[8] *Africa Health Research Institute, Mtubatuba, KwaZulu-Natal, South Africa*
[9] *Centre for Infectious Disease Epidemiology, School of Public Health and Family Medicine, Faculty of Health Sciences, University of Cape Town, South Africa*

Corresponding author: *G Harling (g.harling@ucl.ac.uk)*

Background. South African (SA) national HIV seroprevalence estimates are of crucial policy relevance in the country, and for the worldwide HIV response. However, the most recent nationally representative HIV test survey in 2012 had 22% test non-participation, leaving the potential for substantial bias in current seroprevalence estimates, even after controlling for selection on observed factors.
Objective. To re-estimate national HIV prevalence in SA, controlling for bias due to selection on both observed and unobserved factors in the 2012 SA National HIV Prevalence, Incidence and Behaviour Survey.
Methods. We jointly estimated regression models for consent to test and HIV status in a Heckman-type bivariate probit framework. As selection variable, we used assigned interviewer identity, a variable known to predict consent but highly unlikely to be associated with interviewees' HIV status. From these models, we estimated the HIV status of interviewed participants who did not test.
Results. Of 26 710 interviewed participants who were invited to test for HIV, 21.3% of females and 24.3% of males declined. Interviewer identity was strongly correlated with consent to test for HIV; declining a test was weakly associated with HIV serostatus. Our HIV prevalence estimates were not significantly different from those using standard methods to control for bias due to selection on observed factors: 15.1% (95% confidence interval (CI) 12.1 - 18.6) v. 14.5% (95% CI 12.8 - 16.3) for 15 - 49-year-old males; 23.3% (95% CI 21.7 - 25.8) v. 23.2% (95% CI 21.3 - 25.1) for 15 - 49-year-old females.
Conclusion. The most recent SA HIV prevalence estimates are robust under the strongest available test for selection bias due to missing data. Our findings support the reliability of inferences drawn from such data.

S Afr Med J 2017;107(7):590-594. DOI:10.7196/SAMJ.2017.v107i7.11207

HIV prevalence is possibly the most important indicator for HIV policy. In particular, accurate HIV prevalence estimates are vital for understanding drivers of the worldwide epidemic and planning national resource allocation. Approximately 18% of all people living with HIV in 2013 were estimated to live in South Africa (SA),[1] which had an estimated HIV prevalence of 18.8% among 15 - 49-year-olds in 2012.[2] However, this SA national estimate rests on a survey in which 22% of interviewed individuals declined to take an HIV test. Given this level of non-testing, if everyone who had not tested had in fact been HIV seronegative, seroprevalence would have been 14.7%; if everyone who had not tested had been HIV seropositive, seroprevalence would have been 36.7%.

This wide range of possible HIV prevalence is concerning for several reasons. National HIV prevalence is an important determinant in resource allocation decisions, both internationally and within government and other national service providers. Furthermore, differential bias within the national estimate may lead to misallocation of effort or funds away from higher-risk populations. Finally, changes in biases over time may lead to erroneous conclusions regarding either the effectiveness of programmes or secular trends in infection risk.

National HIV prevalence estimates in SA initially used public sector antenatal testing data,[3] despite awareness that pregnant women attending public sector clinics were not nationally representative.[4] In the early 2000s, SA was among the first countries to undertake a nationally representative population-based HIV survey, and has subsequently performed three more such investigations.[3] There has been increasing concern about response rates in nationally representative surveys potentially allowing for biased HIV estimates.[5] Consent to test among those eligible to be interviewed has been <70% in all four SA surveys.

Missing data in the context of estimating HIV prevalence are only problematic when systematically associated with both HIV status and the unobserved characteristics of those selected for participation ('potential participants'), such that the HIV status of those who are interviewed ('interview participants') but decline to test cannot be predicted from their observed characteristics. If missingness is not systematically associated with HIV status ('missing at random'), or is systematically associated with HIV status but can be predicted from observed characteristics ('missing conditionally at random'), multiple imputation or probability weighting methods will give consistent

estimates of the true population mean. However, if the decision to participate is associated with unmeasured characteristics which are also related to HIV status ('missing not at random') – notably participant knowledge or beliefs about HIV status itself that are unlikely to be observed by those conducting the survey – these methods will not provide valid HIV prevalence estimates. Evidence suggests that non-response to HIV testing surveys can indeed be associated with an individual's HIV status and their knowledge of their HIV status.[6,7]

If non-response is based on unobserved characteristics, Heckman-type selection models can provide consistent HIV prevalence estimates.[8] Selection models make use of variables that predict non-participation but not the outcome of interest – except via its effect on participation. One such variable is the identity of the interviewer who attempted to enrol each potential participant into testing: supervisor-assigned interviewers vary in their ability to persuade invited individuals to participate, but the HIV status of participants is unlikely to be associated with the identity of the interviewer assigned to them.[5,9,10]

Past analyses of HIV prevalence in Africa using interviewer identity as a selection variable have found varying levels of bias, from none to an almost doubling of HIV prevalence.[5,9,11] In SA, a recent analysis of a full-population cohort in rural KwaZulu-Natal Province found significant selection bias.[12] We conducted a selection model analysis on the most recent SA national HIV prevalence survey to determine whether existing estimates are affected by selective survey non-response.

Methods

The 2012 South African National HIV Prevalence, Incidence and Behaviour Survey[2] was a two-stage sample of the SA population, stratified by province, locality type (urban and rural, formal and informal) and race in urban areas. Interviewers were matched to households based on language spoken, and race and ethnicity where possible. An initial household interview was sought with the head of household, after which consent for an individual interview, and subsequently for an HIV test, was sought from each household member. The sample for this analysis comprised all individuals aged ≥15 years living in eligible households who were contacted and consented to an individual interview. We considered including children aged <15 and those with only a household interview, but initial analysis showed that in both groups interviewer identity was only weakly associated with willingness to consent to an HIV test (in children) and to an individual interview (in all ages) (supplementary Table 1) (all supplementary tables and figures are available at http://discovery.ucl.ac.uk/1543319/); we therefore did not consider these groups further. We excluded anyone who was missing age, sex or race information, since these variables were used to match household and individual-level data, as well as anyone with no recorded interviewer identity.

We conducted our selection model analyses using the Semi-ParBIVProbit package in R (version 3.4).[13] This package allows users to implement a range of selection models that extend the original Heckman specification. We jointly estimated a bivariate probit model containing a 'selection' equation to predict consent to HIV testing and an 'outcome' equation to predict HIV status. Both equations contained covariates previously shown to predict either consent or HIV status in a multivariable model.[2] The selection equation also included an indicator for assigned interviewer identity and a ridge penalty term to avoid collinearity in cases where an interviewer's participants all did, or not did not, consent. Effects for continuous covariates were estimated using regression splines, and spatial effects were smoothed at the provincial level using a Markov random field smoother.[14]

We relaxed the typical selection model assumption that the error terms in the two equations are jointly distributed bivariate normal using a copula approach.[14] To ensure model convergence, we restricted ourselves to symmetric Gaussian and Frank copulae as candidate dependence structures, based on preliminary analyses showing that both positive and negative dependence existed in the population (i.e. individuals are both more likely (positive) and less likely (negative) to consent if they believe themselves to be at high risk of being infected).[15] We selected our preferred copula based on the Vuong and Clarke likelihood-based tests.

All models were estimated separately by gender, and we generated separate estimates for 5-year age categories and provinces. Prevalence estimates were adjusted to reflect the national over-15-year-old population using previously generated Human Sciences Research Council (HSRC) household and individual-level questionnaire non-response sampling weights. Confidence intervals (CIs) were calculated from variance-covariance matrices adjusted for stratification and clustering at the first sampling level; however, we did not use weights during model fitting.

Ethical approval

Ethical approval for the original survey was granted by the HSRC's Research Ethics Committee (ref. no. 5/17/11/10). Informed consent was required from each participant.[2] This analysis was exempted from additional review by the Harvard Longwood Medical Area Institutional Review Board (ref. IRB14-4638) because of its use of anonymised existing data.

Results

A total of 42 950 individuals resided in 11 079 participating households (household acceptance rate 89.7%) in the 2012 survey.[2] Matched household and individual questionnaires and valid interviewer identity numbers were available for 42 357 (98.6%) of these, of whom 4 416 (10.4%) were either unreachable or did not consent to interview. Our final analytical sample was 26 708 participants who had completed an individual interview, of whom 6 035 (22.6%) declined to be tested for HIV (supplementary Table 2). Once reweighted, the sample represented 36 699 134 SA residents aged >15 years (i.e. the entire population).

Participants declining an HIV test were more likely to be male, middle-aged (~30 - 50 years (Fig. 1)), white or Asian, Afrikaans or English speaking, married, in the highest wealth household quintile, living in Gauteng and Western Cape provinces, and non-drug users if male and not recent crime victims if female, and to have at least completed secondary education (supplementary Table 3). Non-testing was also higher among those who were older at sexual debut or not yet sexually active, had fewer lifetime partners, thought themselves more likely to become HIV-infected in the future, had tested longer ago if female, and had received their most recent HIV test result if male.

Three hundred and twenty-seven interviewers completed at least one interview with an eligible participant, with varying success in persuading interviewees to test (mean 66%, interquartile range 53 - 85%) (supplementary Fig. 1). Based on Vuong and Clarke tests, we used the Frank copula for males and the Normal copula for females (supplementary Table 4). For both males and females, the association between willingness to test and HIV status was negative overall, more strongly so in the west of the country (supplementary Fig. 2).

HIV prevalence estimates differed little between conventional and selection models (Fig. 2). Our selection models estimated HIV prevalence in males aged 15 - 49 years at 15.1% (95% CI 12.1 - 18.6), compared with 14.5% (95% CI 12.8 - 16.3) in the 2012 HSRC report;[2]

Table 1. HIV prevalence estimates for the 2012 South African National HIV Prevalence, Incidence and Behaviour Survey,[2] stratified by location type

	Sample (unweighted)	Weighted naive		Weighted imputation		Weighted selection model	
		Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Males, ≥15 years							
All	11 564	13.4	12.6 - 14.1	12.3	11.8 - 13.3	14.1	11.3 - 17.5
Urban formal	6 834	11.7	10.8 - 12.6	10.6	9.7 - 11.9	12.7	9.3 - 16.7
Urban informal	1 199	19.6	17.1 - 22.0	18.6	17.0 - 21.0	20.6	16.3 - 25.0
Rural formal	1 300	14.8	12.7 - 16.9	12.9	11.5 - 14.9	14.1	11.7 - 16.9
Rural informal	2 231	13.7	12.1 - 15.2	13.3	12.2 - 15.0	14.8	12.5 - 17.7
Females, ≥15 years							
All	15 144	20.3	19.6 - 21.0	18.8	18.2 - 19.5	19.6	18.3 - 21.9
Urban formal	8 925	15.1	14.2 - 16.0	13.6	12.9 - 14.7	14.9	13.4 - 17.3
Urban informal	1 522	32.5	30.0 - 35.1	31.8	29.5 - 33.6	32.9	31.0 - 36.2
Rural formal	1 284	17.1	14.9 - 19.4	15.7	14.0 - 17.4	16.1	14.4 - 18.4
Rural informal	3 413	24.4	22.9 - 26.0	23.9	22.8 - 25.5	24.3	22.2 - 26.4

CI = confidence interval. CIs allow for stratification and two-stage sampling methodology. Weights were adjustments for non-random sampling and for individuals in the initial sample who were either not contacted or declined to interview. Naive estimates ignore missing HIV outcomes, imputation estimates are based on a model of missing values, and selection estimates are based on a joint model of missingness and HIV status.

selection model HIV prevalence in females aged 15 - 49 years was estimated as 23.3% (95% CI 21.7 - 25.8), compared with 23.2% (95% CI 21.3 - 25.1) in the 2012 report.[2] Smoothed estimates of the associations between HIV prevalence and age, education and wealth suggested little difference between single-equation imputation and bivariate selection models (Fig. 3). When individuals were identified by residential locality type, HIV prevalence remained high among those in urban informal settings, and for females in rural informal settings (Table 1).

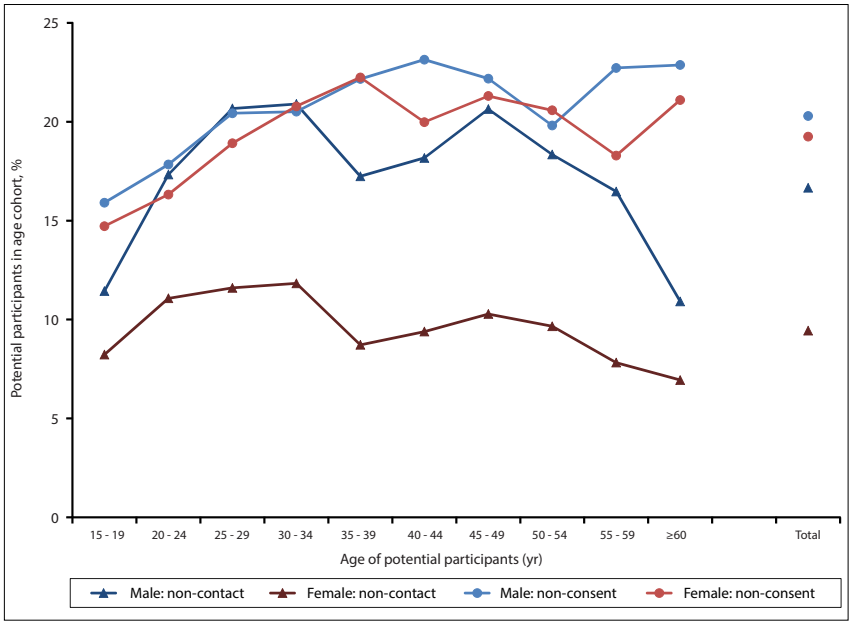

Fig. 1. Age/sex-stratified non-response rates in the 2012 South African National HIV Prevalence, Incidence and Behaviour Survey. ('Non-contact' = potential respondents who were never invited to test for HIV owing to never being interviewed; 'non-consent' = those who were interviewed but declined to test for HIV.)

Discussion

Our analysis suggests that current estimates of SA national HIV prevalence are robust, despite substantial non-participation in HIV testing in the survey from which they are estimated. Using interviewer identity as a selection variable, we found that although interviewers varied in their ability to elicit consent to an HIV test, and although this variation was associated with HIV status, the impact of adjusting for potential selection bias on national estimates of HIV prevalence in SA was small.

Our findings do not preclude the existence of selection effects in SA HIV surveys. One possible explanation for the limited impact of selection in this study is that high-risk populations – notably young women, black Africans and residents of KwaZulu-Natal and Eastern Cape provinces – had the highest consent rates. However, the variability in association between willingness to test and HIV status in SA is of interest, and selection could be relevant for those at both the highest and lowest risk of HIV infection: consent rates were lowest among white and Asian groups, who are at lowest risk, and in Gauteng, where HIV seroprevalence is almost the highest in the country.

A second possible explanation for the similarity between existing estimates and those found in our Heckman-style selection models is that known predictors of HIV status already included in standard methods, such as sociodemographic and behavioural characteristics, also predict the likelihood of testing for HIV and are associated with potential unobserved confounders. Even though many people may be making their decision to test or not test based on information unknown to the interviewer (e.g. they already know their HIV status, or have undertaken unreported risky behaviours that they believe place them at risk of infection), this potential bias in prevalence estimates may already be controlled for if unmeasured characteristics are correlated with factors we are able to adjust for in the model. If this

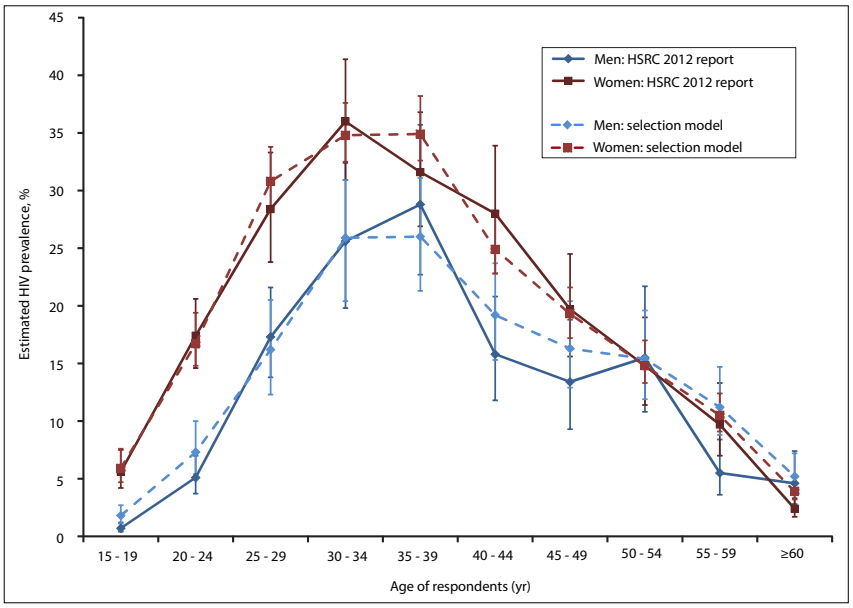

Fig. 2. Comparison of HIV prevalence estimates in the 2012 South African National HIV Prevalence, Incidence and Behaviour Survey[2] and using selection models, stratified by age. Values are prevalence estimates and 95% confidence intervals from sex-specific response-weighted HSRC report[2] and selection models. Numerical values for selection models are provided in supplementary Tables 5 and 6. (HSRC = Human Sciences Research Council.)

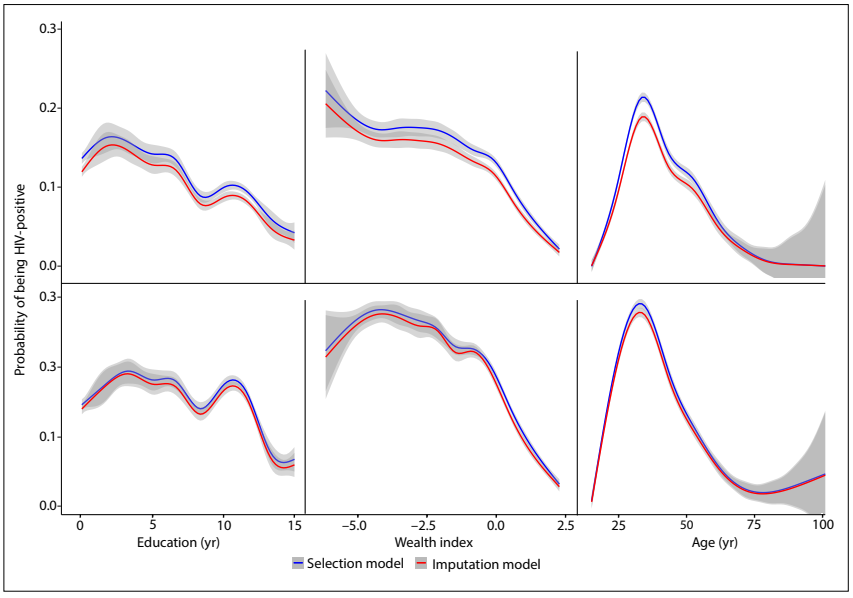

Fig. 3. Predicted HIV prevalence based on multiple imputation and selection models by education, household wealth and age in the 2012 South African National HIV Prevalence, Incidence and Behaviour Survey.[2] Values are smoothed using a spline function; shaded areas represent 95% confidence intervals. (Top row = males aged ≥15 years; bottom row = females aged ≥15 years.)

is the case, unobserved predictors of the likelihood of testing for HIV and of being HIV-positive will be of limited importance, and our methods will not change HIV prevalence estimates. In this scenario, adjusting for observed characteristics in estimating national HIV prevalence will be sufficient to provide valid estimates.

One important change arising from the use of selection models in this study is that we report much wider CIs than those obtained using other methods: our selection model estimate CIs were twice as wide as those from inverse-weighting or multiple-imputation models, and 21% wider for women and 86% wider for men compared with the 2012 HSRC report.[2] This is as expected because non-selection models, by assuming that those who declined an HIV test were missing at random, take the association between consenting to test and HIV status as zero, conditional on covariates. Insofar as this association is not guaranteed to equal zero, could vary, and needs to be estimated, it will introduce greater uncertainty into our estimates.

The strength of our approach is the ability to relax the assumption of missingness at random in estimating HIV prevalence in survey data. By using copulae we were also able to relax the typical selection model assumption of bivariate normal dependence, and we produced subgroup estimates adjusted for selection bias. However, as with any selection model, we rely on the untestable assumption that interviewer identity is a valid selection variable, i.e. is unrelated to HIV status conditional on observed characteristics. Since we did not consider eligible households or individuals who were not were located or declined to interview, we did not capture uncertainty regarding their HIV status, and our confidence bounds should therefore be considered underestimates of uncertainty. While we might expect a weaker association between HIV status and consent to interview than consent to test, a complete assessment of missingness would incorporate all forms of non-contact and non-response.

Conclusion

Our findings are reassuring for ongoing resource allocation, and prevention and care interventions that have been led by existing HIV prevalence estimates in SA.

Acknowledgements. None.

Author contributions. TB and TR conceptualised the study. GH conducted the initial analyses with assistance from MM and SM, and wrote the first draft of the article. GM and MEM provided guidance on the statistical methodologies. All authors contributed to the study design, data interpretation and final revisions to the text.

Funding. The 2012 survey was mainly funded by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention under the terms of Cooperative Agreement Number 3U2GGH000570, with additional financial support from the United Nations Children's Fund, the South African National AIDS Council and the Bill and Melinda Gates Foundation. For this work, GH was supported by the National Institute of Child Health and Human Development (R01-HD084233). TB was supported by the Alexander von Humboldt Foundation through the Alexander von Hum-

boldt professor award, which is funded by the German Federal Ministry of Education and Research; the Wellcome Trust; the European Commission; the Clinton Health Access Initiative; and the National Institutes of Health through the National Institute of Child Health and Human Development (R01-HD084233), the National Institute on Aging (P01-AG041710), the National Institute of Allergy and Infectious Diseases (R01-AI124389 and R01-AI112339) and the Fogarty International Center (D43-TW009775).

Conflicts of interest. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

1. UNAIDS. 2013 Report on the Global AIDS Epidemic. Geneva: UNAIDS, 2013.
2. Shisana O, Rehle T, Simbayi LC, et al. South African National HIV Prevalence, Incidence and Behaviour Survey, 2012. Cape Town: Human Sciences Research Council, 2014.
3. Williams BG, Campbell C. Understanding the epidemic of HIV in South Africa: Analysis of the antenatal clinic survey data. S Afr Med J 1998;88(3):247-251.
4. Zaba B, Boerma T, White R. Monitoring the AIDS epidemic using HIV prevalence data among young women attending antenatal clinics: Prospects and problems. AIDS 2000;14(11):1633-1645. http://dx.doi.org/10.1097/00002030-200007280-00020
5. Hogan DR, Salomon JA, Canning D, Hammitt JK, Zaslavsky AM, Bärnighausen T. National HIV prevalence estimates for sub-Saharan Africa: Controlling selection bias with Heckman-type selection models. Sex Transm Infect 2012;88(Suppl 2):i17-i23. http://dx.doi.org/10.1136/sextrans-2012-050636
6. Korenromp EL, Gouws E, Barrere B. HIV prevalence measurement in household surveys: Is awareness of HIV status complicating the gold standard? AIDS 2013;27(2):285-287. http://dx.doi.org/10.1097/QAD.0b013e32835816ce
7. Reniers G, Eaton J. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. AIDS 2009;23(5):621-629. http://dx.doi.org/10.1097/QAD.0b013e3283269e13
8. Heckman JJ. Sample selection bias as a specification error. Econometrica 1979;47(1):153-161. http://dx.doi.org/10.2307/1912352
9. Janssens W, van der Gaag J, de Wit TFR, Tanović Z. Refusal bias in the estimation of HIV prevalence. Demography 2014;51(3):1131-1157. http://dx.doi.org/10.1007/s13524-014-0290-0
10. Bärnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology 2011;22(1):27-35. http://dx.doi.org/10.1097/EDE.0b013e3181ffa201
11. Clark SJ, Houle B. Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. PLoS One 2014;9(11):e112563. http://dx.doi.org/10.1371/journal.pone.0112563
12. McGovern ME, Marra G, Radice R, Canning D, Newell M-L, Bärnighausen T. Adjusting HIV prevalence estimates for non-participation: An application to demographic surveillance. J Int AIDS Soc 2015;18(1):19954. http://dx.doi.org/10.7448%2FIAS.18.1.19954
13. Marra G, Radice R. SemiParBIVProbit: Semiparametric Bivariate Probit Modelling. R package version 3.4. 2015. https://CRAN.R-project.org/package=SemiParBIVProbit (accessed 1 June 2017).
14. Marra G, Radice R, Bärnighausen T, Wood SN, McGovern ME. A simultaneous equation approach to estimating HIV prevalence with non-ignorable missing responses. J Am Stat Assoc 2016;(epub 26 August). http://dx.doi.org/10.1080/01621459.2016.1224713
15. Radice R, Marra G, Wojtys M. Copula regression spline models for binary outcomes. Stat Comput 2016;26(5):981-995. http://dx.doi.org/10.1007/s11222-015-9581-6

Accepted 22 March 2017.